Bandolier 96

What do we think?
What do we know?
What can we prove?

Evidence-based health care **£3.00** **February 2002 Volume 9 Issue 2**

A rush of blood

There are some topics that are damaging to your health. We try out several this month, in something of a rush of blood to the head. By far and away the most difficult topic is the old chestnut of whether indirect comparisons are valid. If we have tested treatment A against treatment B, and tested treatment C against treatment B, are we able to say anything about the relative merits of A and C?

Some would simply tell us that we could not. Others would suck their teeth like plumbers of old and tell us how tricky it would be even to think about it. A few will just have a go. So ***Bandolier*** this month has a feeble stab at trying to put some sort of sense to it. Highly qualified, because it's so tricky, but there's light at the end of the tunnel.

Not what, but whom

We also have a look at prediction rules, importantly about which men to treat with symptoms of BPH. It all comes down to the size of the prostate and the serum PSA as a surrogate for this. Men who, at randomisation, were given placebo, and who had a serum PSA of 3.3 ng / mL or above, had high rates of spontaneous acute retention (about 8% over four years). They look like the best candidates for treatment, and more complicated algorithms help only a bit better. It's also one reason to measure PSA.

What is marvellous is that this comes from an analysis of randomised trials. The trouble is that it has taken all these years **after** finasteride has been available to get us this information. More please from all those pointy heads in the pharma companies, and soon.

Electronic Bandolier

Just before Christmas ***Bandolier*** Internet had 150,000 visitors in one week. Usually it is a bit lower than that, but it has been growing rapidly in recent months. Much has changed, and for the better. The site is in the final stages of an overhaul to make things easier to find, especially for people using Internet search engines.

In 2002 we have several plans. First is to expand the essays available, and a new one is there on cannabis and flying. This was a spin-off from a planned survey of papers on cannabis and multiple sclerosis that will appear when all papers are in. And in 2002 we plan to develop a resource centre on gout. Few systematic reviews for gout, we have found, though a large Cochrane review is ongoing. We'd like to know what questions you want answered.

CHEMOTHERAPY FOR OLDER PERSONS WITH COLON CANCER?

There are often very good reasons why therapy of older people is different from younger people. For a start, bits drop off as we get older, and we develop more chronic conditions. Older people commonly have to take a lot of medicines. But the number and type of tablets can pose a problem when a new condition arises and new treatment needed. Difficult decisions sometimes have to be made on the balance between efficacy on the one hand and possible harm on the other, let alone which condition is the most important to treat when problems arise.

To some extent that will never go away, but for adjuvant chemotherapy after surgery for colon cancer, we can be reasonably sure that therapy is as effective and no more harmful for patients older than 70 years than those younger than 70 years [1]. The results came from looking at every patient ever entered in a randomised trial and analysing by age.

Review

Randomised trials that randomised patients to chemotherapy after surgery (usually fluorouracil plus calcium folinate) or surgery alone were sought, and authors contacted. Trialists were asked to provide data on outcomes (death, recurrence) for each patient entered, together with toxicity information. Analysis was for overall survival time and recurrence by age of patient.

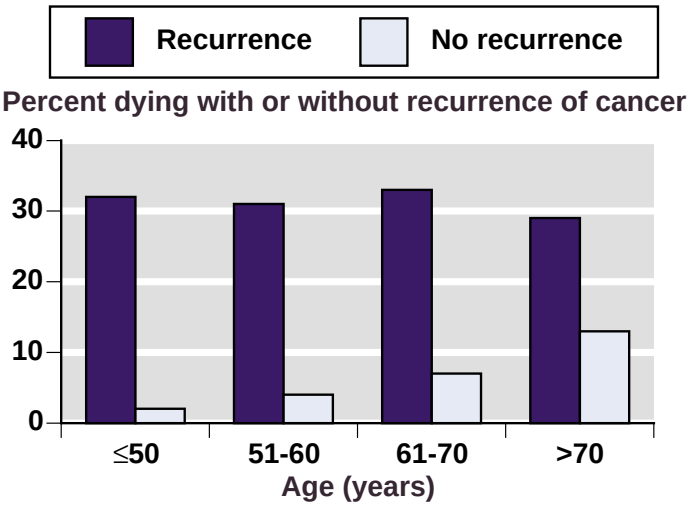
Results

Information was available for seven trials with 2251 patients with stage II or stage III disease. Several trials identified had not completed follow up. Trial size was 239 to 968 patients, with median follow up of five to over eight years, and with six or twelve months of treatment in cycles. Only one trial specified an age limit of less than 75 years.

In this issue

www.ebandolier.com

Figure 1: Percentage of patients with colon cancer dying with recurrence and without recurrence in different age groups



Chemotherapy was effective in increasing survival over five years, from 64% survival in untreated patients to 71% in those treated with adjuvant chemotherapy. The five-year recurrence free rate was increased from 58% in untreated patients to 69% in those treated with adjuvant chemotherapy. The likelihood of this occurring by chance was less than 1 in 1,000 in both cases.

Survival curves for patients older and younger than 70 years were very similar, and age was not important. More patients died without recurrence at older ages (Figure 1), but the proportion dying with recurrence was identical at all ages.

Toxicity was broadly similar for patients less than and more than 70 years. There was no difference in nausea and vomiting, or diarrhoea, or stomatitis, but rates of leucopaenia were about double in patients older than 70 years, and much higher with levamisole (31% at more than 70 years compared with 17% for younger patients) than calcium folinate (8% and 4% respectively).

Comment

Some older people having surgery for colon cancer will not be offered adjuvant chemotherapy for very good reasons. Others may not be offered chemotherapy because of perceptions of greater toxicity, or lower tolerance, or for some other reason relating to age.

This report shows that re-analysis of clinical trial information based on individual patient data can help answer the question of who benefits and who does not. Patients older than 70 years did just as well as younger patients. Important this, as people older than 70 years without cancer have a life expectancy of a decade or so.

References:
1 DJ Sargent et al. A pooled analysis of adjuvant chemotherapy for resected colon cancer in elderly patients. New England Journal of Medicine 2001 345: 1091-1097.

PREDICTORS OF ACUTE URINARY RETENTION IN MEN

Bandolier often hears the complaint that too much attention is given to whether treatments work, and how good they are compared to other treatments, and too little attention is given to help to decide which patient to treat. A possible solution to this problem can come from combined analysis of patients entered into trials, and especially those treated with placebo. If we could determine which of those patients were likely to have an event, then we could treat them to try and avoid it.

Examples are only too rare because pharmaceutical companies, the custodians of the data, often cannot see the need. Praise then when evidence from such analysis looks likely to help decisions about men with clinical BPH [1].

Study

Several large clinical trials have examined the effect of finasteride on men with clinical BPH over two to four years, during which time they were randomised and double-blind. The studies typically excluded men with PSA values above 10 ng/mL, and prostate cancer had to be excluded by biopsy in men with PSA over 4 ng/mL. Men included had moderate symptoms of BPH and a urine flow rate of less than 15 mL/sec.

In all, more than 3,000 men were given placebo, and the rate of spontaneous acute urinary retention (AUR) was known. Data from the men given placebo was used in the analysis. Two thirds of men were randomly selected to form a development set, and the remaining third formed a validation set.

Firstly, 110 baseline clinical variables were identified as potential predictors on the basis of clinical and epidemiological judgement and availability. They included demography, symptoms, bother, urinary flow parameters, comorbidity and concomitant therapy. Prostate volume was available only in a subset of men in one trial.

Logistic regression analysis was used to assess variables in a singly and together in predicting AUR. A potential scoring algorithm was derived using weights from logistic regression coefficients.

Results

On average, men were 64 years old at baseline, had a peak urine flow of 11 mL/sec, and an average symptom score of 15 (moderate), and a mean PSA of 2.9 ng/mL. The spontaneous AUR rate was 2.5% over two years and 3.7% over four years.

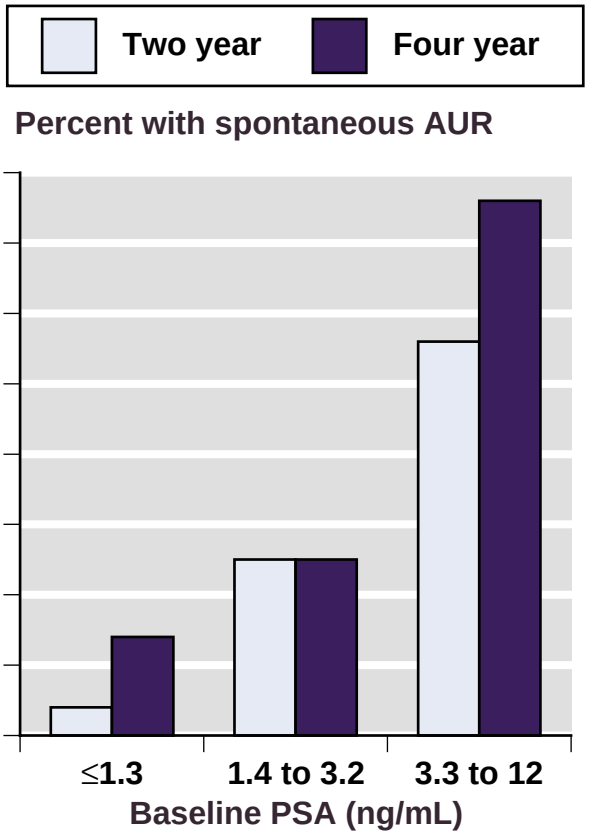
Eventually three systems were chosen for testing. One was a five variable model that included urinating more than every two hours, symptom problem index, maximum urinary flow rate, hesitancy when urinating and PSA. Another was an algorithm using symptom problem index, PSA, uri-

Figure 1: Algorithm for predicting acute urinary retention

Start with serum PSA value

PSA ≥3 ng/mL	AUR likely if peak flow <11.8 mL/sec
PSA<3 ng/mL	AUR likely if: 1 Symptom problem index >8.5, AND 2 Frequent urination few times or less in last month, AND 3 Peak flow less than 8.1 mL/sec

Figure 2: Acute urinary retention in RCTs over two and four years according to initial PSA



nating more often than every two hours and peak urine flow (Figure 1). The third was PSA alone.

Each of these performed about equally well in development and validation sets, and with the combined data (Table 1). If a test performed perfectly in prediction, the area under the ROC curve would be 1, and if it was no better than chance it would be 0.5. Figures of about 0.7 and higher are typical of many tests we use today. PSA alone, for instance, had a sensitivity of 75% and a specificity of 64%, giving a positive likelihood ratio of 2.1.

With PSA alone, the incidence of spontaneous AUR over four and two years was much higher in men in the upper tertile of 3.3 to 12 ng / mL (Figure 2). Over two years at least one man in 20 will have spontaneous AUR if they have symptoms of BPH and a moderately reduced maximum urinary flow rate (less than 15 mL / sec).

Comment

This is useful stuff, though readers should note a couple of things from the paper. Firstly, the decision point for PSA alone is implied as being 3.0 ng / mL, though that is not explicit. Secondly, when the algorithm uses frequent urination, the choice of this being a few times or less in the past month rather than more than a few times looks as if it may be the wrong way round.

The implications, though, are interesting. Since the mean PSA value at baseline in the trials was below 3 ng / mL, it would suggest that perhaps only about half the men who could form a potential treatment group needed to be treated. It would imply that better results of treatment with finasteride might come from men with higher PSA values, though the evidence for this is not available, so far as we know. Perhaps it will form the basis of further analysis of data at the single patient level.

References:
1 CG Roehrborn et al. Clinical predictors of spontaneous acute urinary retention in men with LUTS and clinical BPH: a comprehensive analysis of the pooled placebo groups of several large clinical trials. Urology 2001 58: 210-216.

Table 1: Results for model, algorithm and PSA alone for prediction of AUR

Data set	Number of men	Number of AURs	Area under ROC curve: Five element model	Algorithm	PSA alone
Model development	2146	67	0.71	0.76	0.68
Validation	1016	30	0.74	0.73	0.72
All data	3162	97	0.71	0.75	0.71

HRT AND FRACTURE RISK

Hormone replacement therapy in older women increases bone mineral density, and therefore should help prevent fractures. This is important because nonvertebral fractures occur in 1% of women every year in the decade after their 65th birthday, increasing to 5% a year in women over 85 years (***Bandolier*** 94). Strengthening the skeleton should help. We know that HRT increases bone mineral density. Other treatments that increase bone mineral density reduce fractures. Therefore HRT should reduce fractures. A large case control survey seemed to confirm this (***Bandolier*** 62) with the key message that HRT protects against hip fracture while it is being taken and for a few years afterwards. Continued protection needs continued use.

What happens when we look at two meta-analyses of randomised trials for nonvertebral and vertebral fractures [1,2]? The message is about the same, though perhaps not quite as strong.

Reviews

Both reviews came from the same team at York. Both had exemplary searching, including quizzing folk at international conferences to try and ensure that all published and unpublished trials were found. The key inclusion criteria was that women had been randomised to at least 12 months of treatment, with control of placebo, no treatment, or calcium with or without vitamin D. Most trials reported only mineral density, and fracture data was sought from investigators in most cases.

A sensibly conservative analysis was done with pre-specified sensitivity analyses because of clinical heterogeneity concerning nature of HRT preparation (for instance with or without progestins), dose and duration, as well as nature of women randomised.

Results

Nonvertebral fractures [1]

Twenty-two randomised trials with 556 nonvertebral fractures in 8,776 women were found, eight with unpublished fracture data, and only one setting out to measure fractures as an outcome. Individual trials included 23 to 2,763 women and trial quality was generally good. Duration was 12 to 120 months. There was a wide range of results (Figure 1), with some trials having more fractures with control than HRT (below the line of equality in Figure 1), some having the same, and others having more fractures with HRT than control (above the line of equality). Some trials had no fractures in HRT or control. A data set that some might say was heterogeneous in results.

Using all trials there was a reduced risk with HRT (Table 1), but a significant result was seen only in trials with a mean age of less than 60 years when starting HRT. Pre-specified analysis of placebo-controlled trials showed a significant reduction in risk, as did trials with hip or wrist fractures. Published fracture data showed a significant reduction in fracture risk, while unpublished fracture data did not.

Additional sensitivity analysis with trials with at least one fracture per treatment group, or trials with more than 500 women, also showed a significant reduction in the risk of nonvertebral fractures (Table 1).

Vertebral fractures [2]

Here 13 trials were found, with 98 vertebral fractures in 6,726 women; one study was an abstract. Trial quality was generally good. Duration was 12 to 60 months.

Overall there was a statistically significant reduction in vertebral fractures (Table 1). Most of the effect was in three

Table 1: Summary of results for nonvertebral and vertebral fractures

Data set	Number of trials	Number of women	Fractures (%) with HRT	Fractures (%) with Control	Relative risk (95% CI)	NNT (95%CI)
Non vertebral fractures						
All data	22	8776	5.3	8.0	0.72 (0.56 to 0.93)	36 (26 to 59)
Trials with at least one fracture per group	19	9351	5.5	8.2	0.74 (0.57 to 0.95)	37 (26 to 62)
Trials with at least 500 women	5	5832	6.1	8.4	0.50 (0.36 to 0.71)	43 (27 to 104)
Vertebral fractures						
All data	13	6726	1.3	2.0	0.67 (0.45 to 0.98)	130 (72 to 645)
Trials with at least one fracture per group	8	4692	1.6	2.6	0.64 (0.40 to 1.01)	not calculated
Trials with at least 500 women	2	3766	1.1	1.2	1.05 (0.37 to 2.94)	not calculated

Figure 1: L'Abbé plot of nonvertebral fracture rates for HRT and control. Size of symbol represents the size of the study

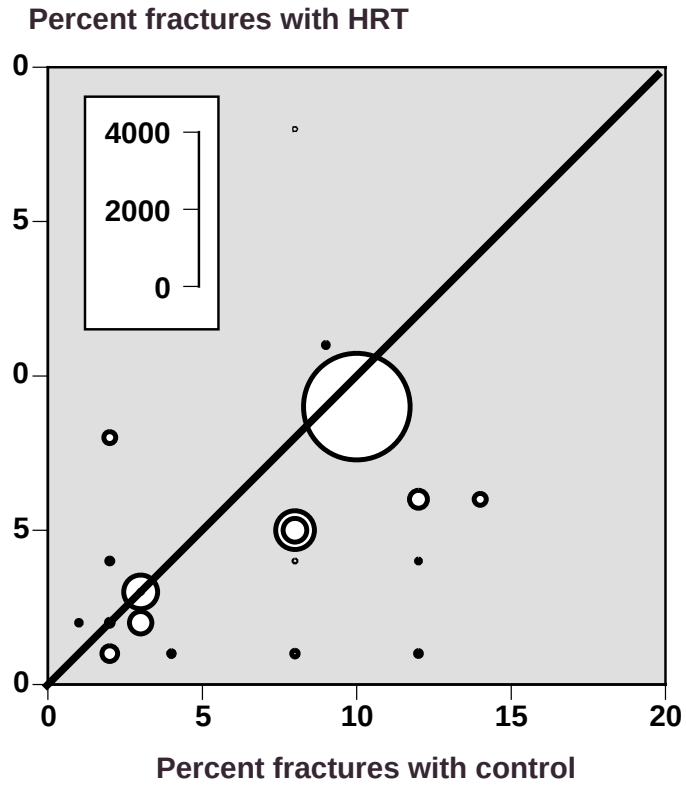
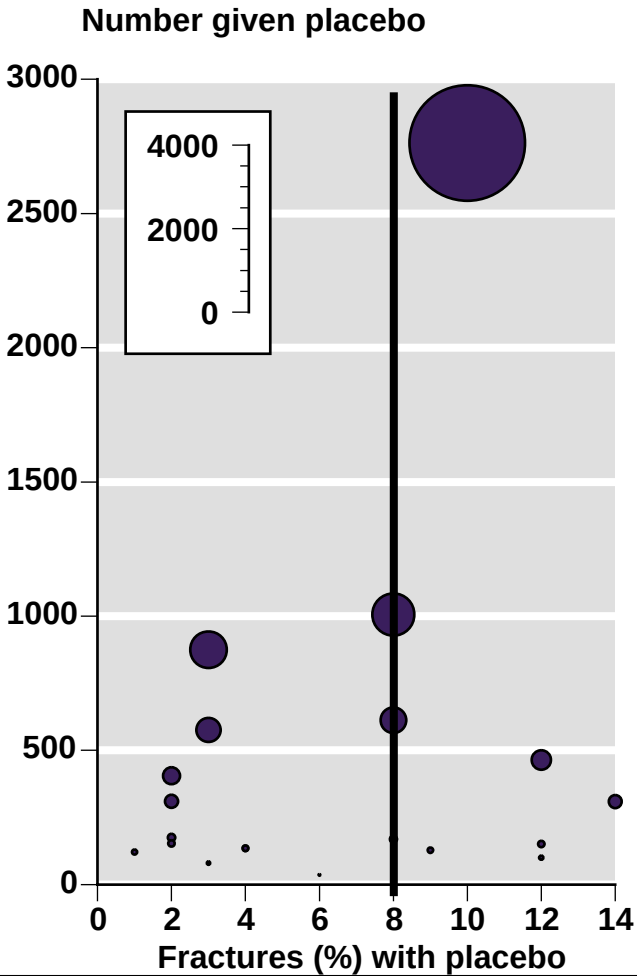


Figure 2: Fracture rate with control (placebo) against number of women given placebo



trials of women with established osteoporosis (though with small numbers), but there was no effect in women without osteoporosis. In five trials with women with a mean age of more than 60 years there was a significant reduction in risk, but not in women starting HRT aged less than 60 years.

Additional sensitivity analysis with trials with at least one fracture per treatment group, or trials with more than 500 women, showed no significant reduction in the risk of nonvertebral fractures (Table 1).

Comment

These are two really excellent systematic reviews tackling an important yet difficult problem. When the rate at which events happen is low, even really effective interventions will need trials with huge numbers to properly demonstrate statistical and clinical significance. This is hardly going to be likely with trials of 23 women, however long the trials.

In the absence of large, carefully designed and executed randomised trials, meta-analysis of small trials is our second best approach. Where there is obvious clinical heterogeneity, as here, we end up pooling information about different (or differing) treatments over different durations. Then trials may be of different quality, and women entering the trials may be different (with or without established osteoporosis, perhaps, or at different ages). The result can be that we get heterogenous results.

But the main problem is size (Figure 2). Low fracture rates in small studies means that event rates are all over the place. Only when the number of women is well over 1000 and we have about 100 fractures can any sense be made of it.

The dilemma is between having sufficient information to pool, and accept a degree of heterogeneity, or salami slice to clinical identity and have too little information to make sense. Sensitivity analysis can help, because it can tell us whether the same order of effect exists (or does not exist) whatever we do. This is one of the best current examples. The authors use some prespecified criteria for sensitivity analysis, and in Table 1 we suggest some others. Together these form a good example for teachers. For those interested in different ways of calculating NNTs, it also shows how clinically heterogeneous trials can sometimes give different results using different methods.

All very academic, but not much use to clinicians making decisions today. What is the answer? Does HRT reduce fractures or doesn't it? On balance it probably does, but the best guess is that about 40 women have to be treated for one to ten years to prevent a fracture in one of them.

References:

1 DJ Torgerson & SE Bell-Syers. Hormone replacement therapy and prevention of nonvertebral fractures. JAMA 2001 285: 2891-2897.

2 DJ Torgerson & SE Bell-Syers. Hormone replacement therapy and prevention of vertebral fractures: a meta-analysis or randomised trials. BMC Musculoskeletal Disorders 2001 2: 7. (www.biomedcentral.com / 1471-2474 / 2 / 7.)

MINDSTRETCHER – BEING INDIRECT

One of the toughest tasks we face is evaluating the place of new treatments. It is relatively straightforward when we have no current treatment and a new one comes along. But when we have several treatments and we want to know how a new one compares, then it is not so easy. What we would like, of course, is large, randomised, valid, head-to-head comparisons. There are almost never available, but what we have is a series of trials with different comparators. What can we do with this?

Indirect method?

Back in 1997 a group from McMaster came up with a method that allowed the calculation of odds ratios or relative risk of A versus B when we have only A versus C and B versus C trials [1]. Essentially it takes the ratio of the log odds of A versus C and B versus C studies.

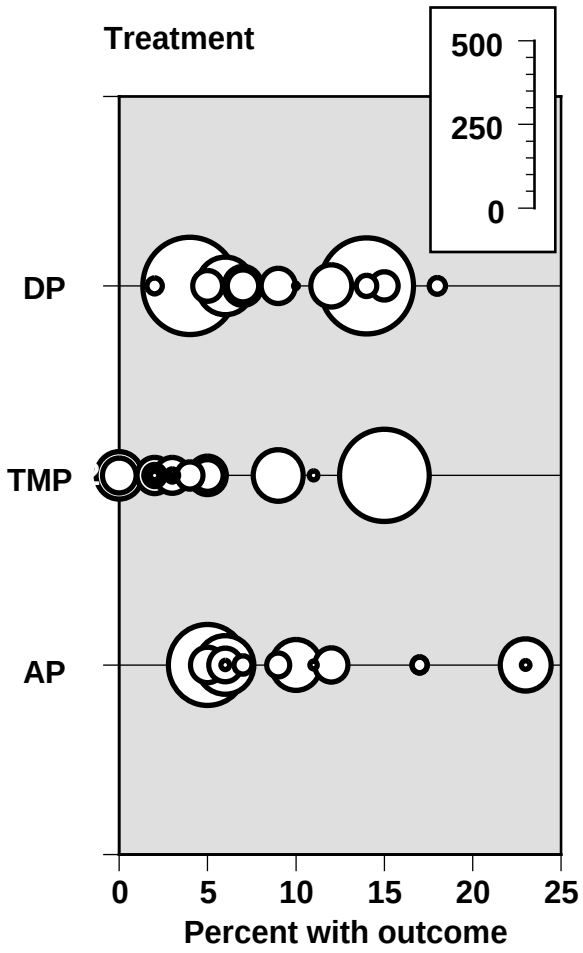
There are lots of equations, and it is not easy to get a simple brain around it. Even though it looks sensible, much still depends on the data sets to which methods might be applied. Statistics can't rescue us from inadequate or insufficient evidence. So three examples to look at how we might expand our thinking on indirect comparisons.

1 P carinii pneumonia [1]

This was the original data set on which the indirect calculation of odds ratios was based. The setting was a systematic review of antibiotic regimens for the prevention of P carinii pneumonia in patients with HIV infection. There were two experimental therapies, trimethoprim-sulphamethoxazole (TMP) and dapsone / pyrimethamine (DP) and a standard therapy of aerosolised pentamidine (AP).

Results of the trials in direct comparisons is shown in Table 1. TMP was better than AP, DP was no different from AP, and TMP was better than DP. Odds ratios calculated using the new method from indirect comparisons was close to those from direct trials. Figure 1 shows an abacus plot of the single treatment arms for TMP, DP and AP. Overall, P carinii pneumonia occurred in 5.5% (95% CI 4.4 to 6.7%) of

Figure 1: Percent with P Carinii pneumonia with AP, TMP and DP in direct and indirect randomised trials



1484 patients taking TMP, 9.0% (7.6 to 10.4%) of 1547 patients taking DP and 9.9% (8.3 to 11.5%) of 1331 patients taking AP.

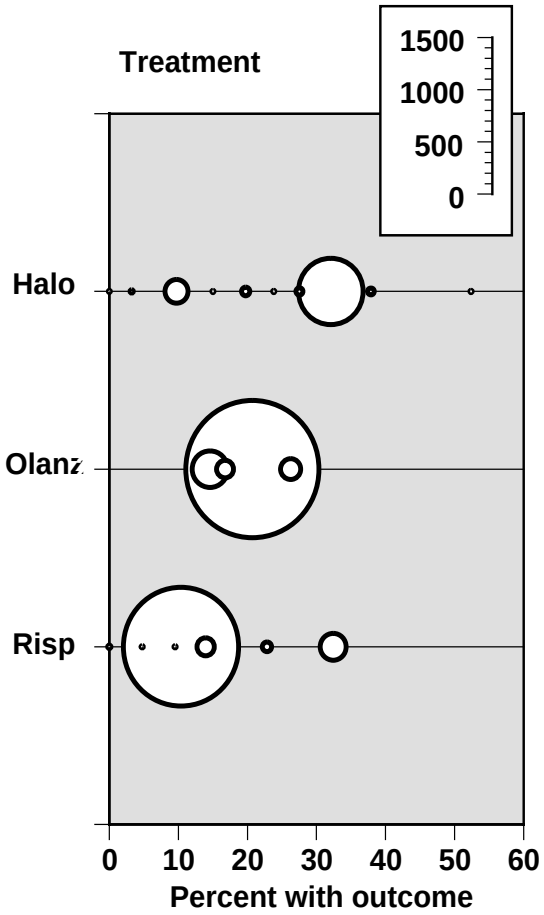
The authors rightly spent some time worrying about whether direct comparison trials were different from other trials, because, for instance, they were longer. Most trials were small and had few events, some none. The indirect comparison, the direct comparison, and a simple graphical representation of the trial arm data all gave the same overall conclusion, that on this measure of efficacy TMP was the clear winner.

Table 1: Information on P carinii pneumonia in direct and indirect randomised comparisons of TMP, DP and AP

Comparison	Number of trials	Number/total (%) with outcome		Relative risk	NNT
		Treatment	Comparator		
TMP vs AP	9	26/681 (4.0)	74/613 (12.3)	0.35 (0.23 to 0.53)	12 (9 to 19)
DP vs AP	5	51/732 (7.0)	58/718 (8.1)	0.89 (0.62 to 1.27)	NA
TMP vs DP	8	56/803 (7.1)	88/815 (10.9)	0.66 (0.48 to 0.90)	26 (15 to 96)

N/A = not applicable

Figure 2: Percent discontinued because of lack of efficacy in randomised comparisons of risperidone, olanzapine and haloperidol



2 Newer antipsychotics for schizophrenia [2]

This topic is a great deal more difficult. For a start, efficacy in schizophrenia trials is measured using several different scales, few of which make much sense or are interpretable for everyday clinical practice. In consequence, outcomes like discontinuation (for adverse events or lack of efficacy) is often the most useful measure. Then there is the question of dose or older or newer antipsychotics. Doses are often titrated, but trials may used fixed doses, some of which are less effective than older antipsychotics used at fixed dose. And there are many nuances that make this even more complicated.

Anyway, the issue here was between risperidone and olanzapine, both of which had been compared with haloperidol in a number of trials, and for which there was one direct comparison. Table 2 shows the data for withdrawal due to lack of efficacy.

In the direct comparison for this outcome, risperidone failed to beat haloperidol, while olanzapine did beat it. There was no difference in the direct comparison. A difficulty was that the rate of discontinuations for lack of efficacy with haloperidol in the olanzapine trials was quite a lot higher than in the risperidone trials.

An abacus plot of data from all treatment arms (Figure 2) emphasises the dependency on some large trials. Overall, lack of efficacy withdrawal occurred in 14% (13 to 16%) of 1745 patients on risperidone, 20% (18 to 22%) of 2027 patients on olanzapine and 26% (23 to 28%) of patients on haloperidol.

The review [2] concluded that over all outcomes the indirect and direct comparisons gave the same answer. That is probably correct. Is olanzapine better than risperidone? That's more difficult, though if for this example anticholinergic drug use had been chosen as an outcome, it would have shown more use with risperidone than olanzapine.

3 Paracetamol and codeine [3]

Now a subtly different examination of indirect comparisons, using the combination of paracetamol 1000 mg plus codeine 60 mg. This combination has been shown to have a low (good) NNT in standard high-quality acute pain trials, but we have only three trials with under 200 patients, only 114 of whom had paracetamol and codeine. How can this result be buttressed?

One way is to look at other paracetamol / codeine combinations in a similar setting (Table 3). Demonstrating a dose-response relationship is helpful, and we get a similar dose-response if we look at an abacus plot (Figure 3) from all placebo controlled trials.

Other things that could be done would include assessing how accurate we can be with this level of efficacy and amount of data (92% confident that we are within ±0.5 of the true NNT). We'd need data from 100 more patients to be at least 95% confident.

Table 2: Discontinuation because of lack of efficacy in randomised comparisons of risperidone, olanzapine and haloperidol

Comparison	Number of trials	Number/total (%) with outcome		Relative risk	NNT
		Treatment	Comparator		
Risperidone v haloperidol	7	225/1573 (14)	68/404 (17)	0.81 (0.63 to 1.03)	NA
Olanzapine v haloperidol	3	375/1860 (20)	239/786 (30)	0.68 (0.59 to 0.78)	10 (7 to 15)
Risperidone v olanzapine	1	24/172 (14)	28/167 (17)	0.83 (0.50 to 1.37)	NA

N/A = not applicable

Table 3: Comparisons of different combinations of paracetamol and codeine with placebo in acute pain studies, with outcome of at least 50% pain relief over 4-6 hours

Comparison	Number of trials	Number/total (%) with outcome		Relative risk	NNT
		Treatment	Comparator		
1000 mg + 60 mg	3	65/114 (57)	9/83 (11)	4.8 (2.6 to 8.8)	2.2 (1.7 to 2.9)
600/650 mg + 60 mg	13	191/398 (48)	78/418 (19)	2.5 (2.0 to 3.1)	3.4 (2.8 to 4.3)
300 mg + 30 mg	4	56/215 (26)	14/164 (9)	3.2 (1.8 to 5.6)	5.6 (4.0 to 9.8)

In three trials with active controls there was information on another 117 patients given paracetamol / codeine at the doses interesting to us, and they had a similar rate of pain relief. In another six trials omitted from the meta-analysis because of technical problems with measurement scales rather than design issues that could lead to bias, the combination of paracetamol and codeine was better than placebo or comparator on at least one measure.

So the sparse data in indirect studies can be supplemented with considerable amounts of information from other high-quality trials.

Comment

Indirect comparisons make for the biggest problems and arguments. The bottom line is that there is no doubt but that best information will come from large, properly constructed randomised trials, using valid outcomes, and done in a way that is meaningful to clinical practice. In most cases this is nothing more than baying for the moon.

When we need to make decisions now based on the information we have, we will be forced to look at indirect comparisons. The simple rule is that quality cannot be compromised. Data from trials prone to bias because of faulty design won't help us, and may drive us to an incorrect conclusion. Then we have to use outcomes that make sense. And we need sufficient numbers of patients and events to overcome any random effects.

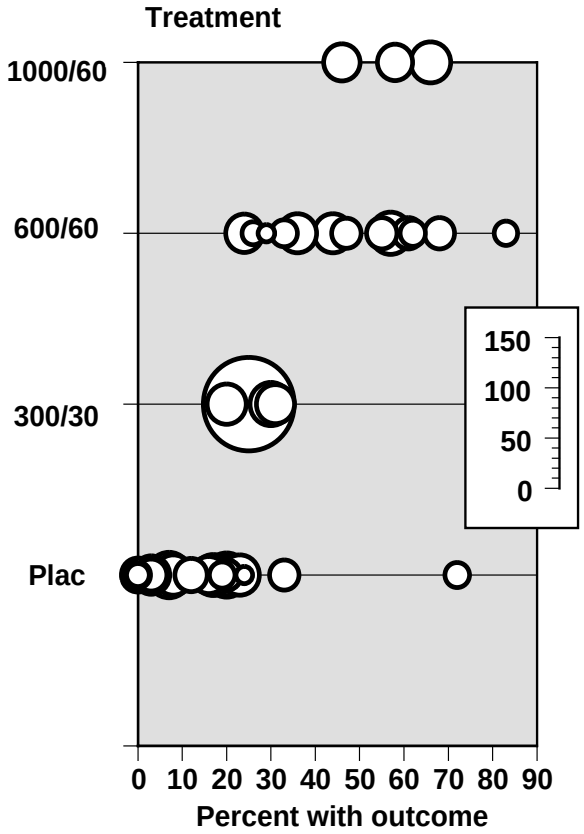
After that, we're probably on our own, though indirect odds ratio calculations may be useful [1] in some circumstances. Abacus plots of single trial arms can be useful back-ups, but have the problem of losing the advantage of randomisation unless there is excellent clinical homogeneity to begin with (and even then use them with caution until we know more).

References:

1 HC Bucher et al. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997 50: 683-691.

2 L Sauriol et al. Meta-analysis comparing newer antipsychotic drugs for the treatment of schizophrenia: evaluating the indirect approach. Clinical therapeutics 2001 23: 942-956.

Figure 3: Percent with outcome of at least 50% pain relief over 4-6 hours for placebo, paracetamol 300 + codeine 30, paracetamol 600 + codeine 60 and paracetamol 1000 + codeine 60



3 LA Smith et al. Using evidence from different sources: an example using paracetamol 1000 mg plus codeine 60 mg. BMC Medical Research Methodology 2001 1:1 (www.biomedcentral.com / 1471-2288 / 1 / 1).

EDITORS

Andrew Moore Henry McQuay
Pain Relief Unit
The Churchill, Oxford OX3 7LJ

Editorial office: 01865 226132
Editorial fax: 01865 226978
Email: andrew.moore@pru.ox.ac.uk
Internet: www.ebandolier.com
ISSN 1353-9906